AIPI 590 PROJECT 1

Keese Phillips

Contents

Discussion

Dataset used

The dataset used to train the model is the Scene UNderstanding (SUN) Database. The dataset was developed by J. Xiao, et. al. as discussed within their paper, *SUN Database: Large-scale Scene Recognition from Abbey to Zoo* (J. Xiao, et. al.). The dataset is comprised of 108,754 images of 397 classes for scene recognition. Although the original paper contains 908 classes, the 397 classes serve as a sufficient comparison for various state-of-the-art (SOTA) models. The smaller 397 class dataset serves as the dataset used to train the model discussed in this report.

Prior work

Feifei Lee et. al. discussed in *Scene recognition: A comprehensive survey the various SOTA recognition algorithms* (Feifei Lee et. al.). The paper discusses the transformations that occurred in scene recognition over the years, starting with Global Attribute Descriptors in the early 2000s, these algorithms were largely limited by their ability to determine complex visual constitutions. Next the field moved toward Patch Feature Encoding, where the model extracts local features from image patches and then approximates these features into a codebook of visual words. The model represents the entire image as a histogram of these visual words, while also attempting to preserve some spatial information. Spatial Layout Pattern Learning then used the underlying spatial layouts to promote the recognition performance. Discriminative Region Detection autonomously selects the crucial regions for scene recognition (Feifei Lee et. al). Object Correlation Analysis “seeks to model the relations between the distribution of diverse objects and scene categories. Early explorations of object correlation are built on topic models in which object recognition is a prerequisite” (Feifei Lee et. al.). Hybrid Deep Models proved to have become one of the most effective methods for scene recognition by harnessing the benefits of multiple methodologies.

Block diagram

See Appendix I

Methodology

The SUN Database needed to be transformed prior to training the model, this enabled the model to become more generalizable. The images were also augmented randomly, with 0 to 2 of the augmentations applied to each image, these image augmentations further allow the model to become more robust and generalizable. The augmentations include horizontal flips with a probability of 50%, which can help the model to recognize objects regardless of orientation, random cropping up to 10% of the image, which can simulate zoom effects and ensure the model can focus on various parts of an image, rotations between -20 and 20 degrees, which can allow the model to adapt to slight changes in orientation, and a Gaussian blur with a sigma range of 0 to 3.0, which can help to reduce noise and focus on more significant patterns. To create a model for scene recognition, transfer learning was employed. Specifically, the ResNet152 model was used. The weights were frozen from the ImageNet1K version 2 model, and the classification layer was removed. The ResNet152 model must be given images that have been cropped, resized, and normalized prior to training. An average pool layer was added, then a linear layer, ReLU layer, Dropout layer, and then a final linear layer for classification. Cross Entropy Loss was the loss function used since the problem in nature is a multi-class classification issue.

Analysis of results

Comparison

Lopez-Cifuentesa et. al. employed a "novel strategy" to detect scenes within the SUN Dataset. They used object level information to decipher the scene during the learning process, this seems to resemble the spatial layout pattern learning as previously mentioned by Feifei Lee et. al. (Lopez-Cifuentesa et. al.). With this methodology, Lopez-Cifuentesa et. al. achieved a 74.04% top-1 accuracy. Their model outperformed all other models on the SUN Databasein terms of accuracy, see Appendix II, as described by *Scene recognition with objectness* (Cheng et. al). The traditional methods when tested on the SUN Dataset achieved a top-1 accuracy of between roughly 28-40%. The CNN approaches achieved a top-1 accuracy between roughly 40-74%. The ResNet152 fine-tuned model achieved a 59.18% and 85.84% top-1 and top-5 accuracy respectfully. This model outperformed all of the traditional approaches and several CNN-based approaches but failed to outperform the top-1 accuracy set by Lopez-Cifuentesa et. al. of roughly 74% (Lopez-Cifuentesa et. al.).

Evaluation

When training the model, a consideration was the possibility of overfitting the training set. The ResNet152 model has close to 60 million parameters, of which roughly 98% would have their weights frozen during the training process. Therefore, the total amount of trainable parameters is small in comparison to the entire model. A smaller number of epochs was chosen so as to not overfit the training set, see Appendix III, while the validation accuracy does not significantly decrease, after 3 epochs the validation accuracy flatlines to roughly 59%. In this case, the model does not seemingly benefit from the final two epochs of training, although it does not readily appear the model has overfit the training set. In terms of individual class accuracy, the fine-tuned

ResNet152 model predicted the classes with the highest and lowest accuracy roughly in line with the original paper as seen in Appendix IV, for the classes with the highest accuracy, and Appendix V, for the classes with lowest accuracy (Xiao, Jianxiong, et al.).

References

Cheng, Xiaojuan, et al. "Scene recognition with objectness." *Pattern Recognition* 74 (2018): 474-487.

Lopez-Cifuentesa, Alejandro, et al. *Semantic-Aware Scene Recognition*, Universidad Autonoma de Madrid, 22 Jan. 2020.

Xiao, Jianxiong, et al. "Sun database: Large-scale scene recognition from abbey to zoo." *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010.

Xie, Lin, et al. Scene Recognition: A Comprehensive Survey, Pattern Recognition, 15 Jan. 2020, www.sciencedirect.com/science/article/abs/pii/S003132032030011X.

Appendices

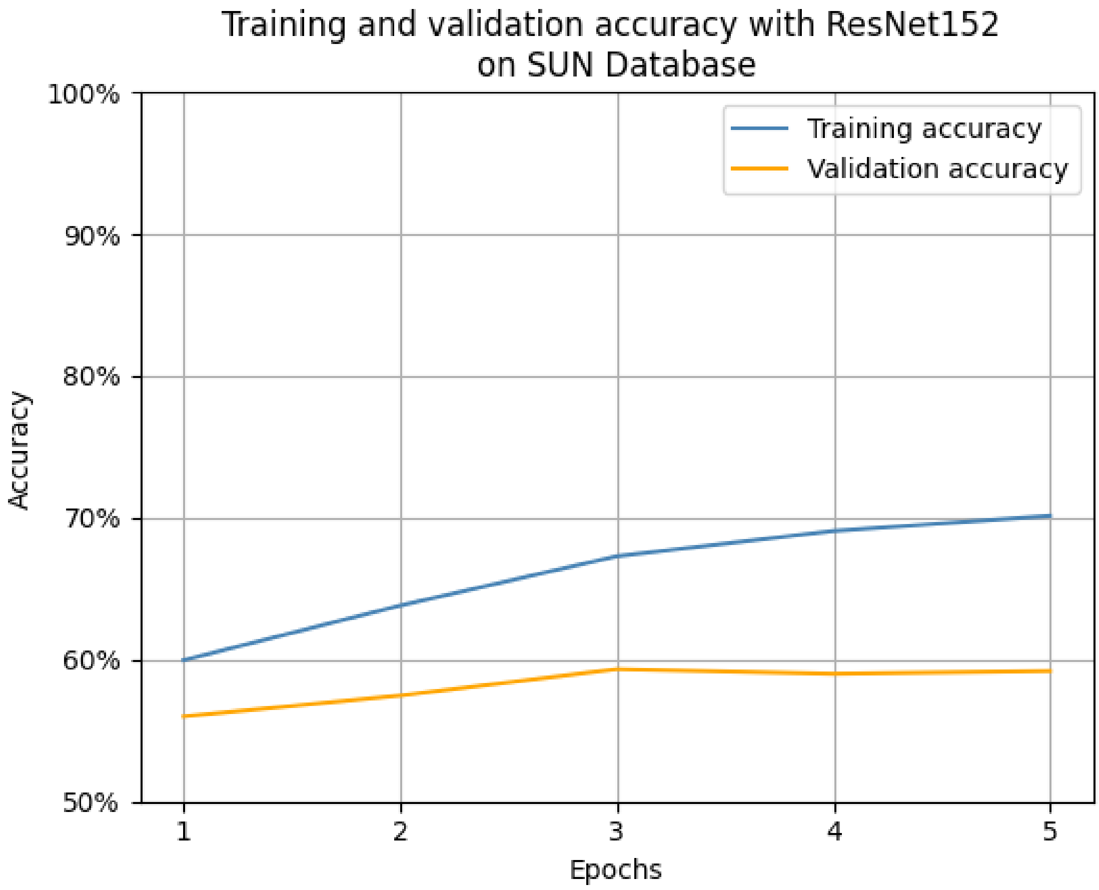
Appendix I

https://github.com/keesephillips/applied_computer_vision_project_1/blob/main/assets/block_diagram.png

Appendix II

Traditional Methods	**Accuracy (%)**
S-manifold	28.90
OTC	34.56
contextBoW + semantic	35.60
Xiao *et al*	38.00
FV (SIFT + Local Color Statistic)	47.20
OTC + HOG2 × 2	49.60
CNN based Methods	**Accuracy (%)**
Decaf	40.94
MOP-CNN	51.98
HybridNet	53.86
Places-CNN	54.23
Places-CNN ft	56.20
CS(VGG-19)	64.53
VSAD	73.00

Appendix III



Appendix IV

Class	**Accuracy**
wine_cellar/barrel_storage	0.920000
oilrig	0.928571
ball_pit	0.951220
florist_shop/indoor	0.967742
underwater/coral_reef	0.969697

Appendix V

Class	**Accuracy**
biology_laboratory	0.0
bistro/indoor	0.0
hunting_lodge/outdoor	0.0
inn/outdoor	0.0
synagogue/indoor	0.0